Croatian Journal of Philosophy
Vol. XVII, No. 52, 2018

"The Brain in Vat" at the Intersection

DANILO ŠUSTER
University of Maribor, Maribor, Slovenia

Goldberg 2016 is a collection of papers dedicated to Putnam's (1981) brain in a vat ('BIV') scenario. The collection divides into three parts, though the issues are inter-connected. Putnam uses conceptual tools from philosophy of language in order to establish theses in epistemology and metaphysics. Putnam's BIV is considered a contemporary version of Descartes's skeptical argument of the Evil Genius, but I argue that deception (the possibility of having massively false belief) is not essential, externalism does all the anti-skeptical work. The largest section in the collection covers Putnam's model-theoretic argument (MTA) against metaphysical realism (MR) and its connections with the brain in vat argument (BVA). There are two camps—unifiers (there is a deep connection in Putnam's thoughts on BVA, MTA and MR) and patchwork theorists and I try to provide some support for the second camp. All of the papers in the collection are discussed and the anti-skeptical potential of BVA is critically assessed.

Keywords: Putnam, brain-in-a-vat scenario, skepticism, realism, model-theoretic argument.

It is not easy to track the provenance of the *brain in a vat* ('BIV' for short) scenario. The contemporary empirical source seems to be the work of Canadian neurosurgeon Wilder Graves Penfield on neural stimulations (in the 1930s) and experiments in waking human subjects undergoing epilepsy surgery. Penfield observed quite complex memories being switched on by electrical stimulation of the appropriate parts of the cerebral cortex (Tallis 2011: 36). Its philosophical use is (first?) registered in the work of Gilbert Harman (1973)—a playful brain surgeon might be giving you "normal" experiences by stimulating your cortex in a special way, but in reality "you might really be stretched out on a table in his laboratory with wires running into your head from a large computer. Perhaps you have always been on that table. ... Or perhaps you do not even have a body. Maybe you were in an accident and all that could be saved was your brain, which is kept alive in the laboratory" (Harman 1973: 5). This type of scenario leads to

familiar philosophical problems of other minds and the external world skepticism, evoked, famously by Descartes. Recall: "... some evil spirit, supremely powerful and cunning, has devoted all his efforts to deceiving me. ... What truth then is left? Perhaps this alone, that nothing is certain" (Descartes 2008: 16).

Nowadays the scenario is almost automatically associated with Hilary Putnam (the first chapter of his 1981). An entire new collection (Goldberg 2016 in the series on *Classic Philosophical Arguments*) is now dedicated solely to philosophical applications and ramifications of the version proposed by Putnam. Descartes is still in the background, thus *Goldberg* in *Introduction* (2016: 2) "Putnam's reflections on the BIV scenario have a familiar historical precedent, of course, in Descartes's reflections on the Evil Demon scenario." The connection with the Cartesian deceiver is not entirely accurate and I find the proper role of deception to be controversial. Putnam actually writes: "Perhaps there is *no* evil scientist, perhaps (though this is absurd) the universe just happens to consist of automatic machinery tending a vat full of brains and nervous systems" (Putnam 1981: 6). In Putnam's BIV world everyone is raised as brains in vats, but their perceptual input is qualitatively just like ours. Could this be our predicament? Putnam argues from some plausible assumptions about the nature of reference to the conclusion that it is *not* possible that all sentient creatures are brains in a vat. A deceivingly simple and enormously influential argument ('BVA' for short) in various fields of philosophy. The collection divides into three parts, though the issues are inter-connected. Putnam uses conceptual tools from philosophy of language in order to establish theses in epistemology and metaphysics.

The first part, "Intentionality and the philosophy of mind and language" opens with an essay by Anthony *Brueckner*, one of the earliest commentators who wrote several papers on the argument. His seminal paper reconstructed the argument in terms of a disjunctive dilemma suggested by Putnam (Brueckner 1986: 154; more or less reproduced by *Pritchard and Ranalli* in Goldberg 2016: 78):

(1) Either I am a BIV (speaking vat-English) or I am a non-BIV (speaking English).
(2) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are true iff I have sense impressions as of being a BIV.
(3) If I am a BIV (speaking vat-English), then I do not have sense impressions as of being a BIV.
(4) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are false. [(2), (3)]
(5) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are true iff I am a BIV.
(6) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are false. [(5)]
(7) My utterances of 'I am a BIV' are false. [(1), (4), (6)]

Whatever proposition is expressed by my utterances of 'I am a BIV' is a false proposition. The anti-skeptical conclusion seems to be that I therefore know that I am not a BIV. The argument is based on an analysis of the truth conditions for the sentences uttered (or thought) by a BIV. These conditions depend on the assignments of references which one would make in evaluating the truth value of BIV's utterances. According to semantic *externalism* when S uses a referring term, she refers to whatever typically causes her uses of that term (in the case of BIV—sense impressions as of being a BIV, according to Brueckner and many other commentators, but not real "brains" and "vats").

The exact role and type of *externalism* used in the argument has been disputed, however. *Kallestrup* (Goldberg 2016: 53) argues that the causal constraint on reference needed in Putnam's proof is actually quite weak and consistent with semantic internalism: "semantic externalists are no better placed than semantic internalists in terms of being able to appeal to Putnam's proof as a semantic response to epistemological skepticism." *Grundmann* (Goldberg 2016: 90–110) on the other hand compares the New Evil Demon (NED) intuition—one can have justified beliefs about the world even if one is living in a demon world with the Old Evil Demon (OED) intuition (BIV, dream). According to the latter one cannot possess justified beliefs about the world unless one is able to rule out relevant skeptical hypotheses. There was always a strong tendency to regard the NED intuition as evidence for the internalism, but Grudmann argues that the NED intuition does not provide a compelling argument for mentalism but is in fact compatible with the view that justification requires reliability. The BVA assumes the view that the individuation conditions of mental content depend, in part, on external or relational properties of the subject's environment. If these connections are constructed reliabilistically and reliability is a necessary condition for justification this would vindicate the crucial role of externalism in Putnam's argument, or so it seems.

An interesting new development in this area is explored by *Bernecker* (Goldberg 2016: 54–72). Whereas content externalism locates mental states inside the head or body of an individual, the hypothesis of *extended mind* claims that the role of the physical or social environment is not restricted to the determination of mental content. Mental states are not only externally individuated but also externally located states. Just as the brain in a vat forms a coupled system with the supercomputer that feeds it all of its sensory-input signals, the supercomputer forms a coupled system with the evil scientist who programs it (Goldberg, ed. 2016: 64). But the scientist presumably speaks a "normally" referring language, and since the brain in a vat should count as an extension of the evil scientist's mind it too, can, after all refer to trees and vats and so on. When content externalism is combined with the extended mind hypothesis it is robbed of its anti-skeptical power according to Bernecker.

The topic of externalism, self-knowledge and reliabilism in the form of sensitivity principle is also discussed by *Becker* (Goldberg 2016: 111–127). The crucial belief "I am a not BIV" is *sensitive* (and thus fulfills a necessary condition for knowledge), for if it were false I would not believe that I am. "I would have some other belief, such as that I am not some specific state type of some particular automated machinery" (Goldberg 2016: 116). But unless I *know* that my terms are referring and my thoughts are about brains and vats, I don't know whether the belief that I express by 'I am not a BIV' is that I am not a BIV. The appeal to sensitivity has not explained how I could know that the skeptical hypothesis is false. Becker's result is largely negative—sensitivity adds nothing to the standard view and standard discussion.

Standard discussion views the BIV scenario primarily as a vehicle for Cartesian angst (cf. *Button* in Goldberg 2016: 142). The worry that it generates is that appearances might be radically *deceptive*, so that (almost) all of our beliefs are *false*. In particular, my utterances of 'I am a BIV' are false if I am a BIV (speaking vat-English), according to Brueckner (recall step 4 in the disjunctive argument above). The vat-English truth conditions of 'I am a BIV' are not satisfied because of *deception* (I am not fed experiences about my "reality", representing me to be a disembodied BIV). But I think that deception, implying *false* beliefs, is, strictly speaking, not essential at all. On the assumption of externalism BIVs lack conceptual resources to even think about the reality of their situation. The Evil Demon scenario has undergone an important historical transformation.

We should follow the suggestion by Miščević (Miščević 2016) and explore the diachronic developments in a long-term life of a thought experiment. The BIV scenario lies at the intersection of "trails" of two thought experiments, the Cartesian Evil Demon scenario and Putnam's *Twin Earth* scenario (Oscar on the Twin Earth, not being in causal contact with Earthly H_2O, does not refer to water). Deception is of course crucial in the Cartesian scenario, but when the two scenarios are combined all the anti-skeptical work is done by semantic externalism—in order for our words to refer to a particular kind of thing, it is necessary for our uses of the term to be connected in an appropriate way with things of that kind. Recall Putnam's initial analogy: an ant is crawling on a patch of sand and as it crawls, it traces a line in the sand which ends up looking like a caricature of Winston Churchill (Putnam 1981: 1). The Putnamian intuition is that the caricature does not refer to or represent Churchill, because the *presuppositions* of successful reference are not fulfilled. This suggests that the main problem with BIV mental states is not a cruel deception, but lack of proper connection.

Suppose we take seriously the parenthetical part of Putnam's own comment (Putnam 1981: 15): "the sentence 'we are brains-in-a-vat' says something false (if it says anything)." We should then reconsider the anti-skeptical argument not on the assumption that "We are not brains in a vat" is false, rather, the preconditions for its being true or false

are not fulfilled (I try to do this in Šuster 2016). To repeat, I think that Putnam's externalism is the basis of his reply to BIV skepticism: no false beliefs because no real beliefs (thoughts) at all (and not because some demonic machinery is feeding us *false* impressions). Still, a vast majority of authors in the collection take the crucial role of massively false beliefs for granted (with *Folina* as an exception).

I will return to the assessment of the Putnam-style refutation of radical skepticism later (Part II: "Epistemology"). Let me jump to the third and the largest section, "Metaphysics", covering Putnam's model-theoretic argument (*MTA*) against metaphysical realism (MR) and its connections with the brain in vat argument (*BVA*). It is a vexed issue how to reconstruct interrelations between MTA, BVA and MR. Even Putnam himself is (characteristically) ambiguous. According to his own report (Putnam 1992: 362):

> I gave a seminar at Princeton in the late seventies at which I presented and defended my model-theoretic arguments. David Lewis, who was present, commented that "there must be something wrong somewhere"—because, if my arguments were right, it followed that we could not be brains in a vat!

So there is a direct connection between the BIV scenario and the model-theoretic argument, MTA implies BVA? But there are other reports, for instance by Brueckner, who thinks that BVA should be sharply distinguished from the model-theoretic argument against metaphysical realism (1986: 149, footnote 2):

> Putnam has indicated (in conversation) that it was in fact his intention to construct an argument in chapter 1 [of Putnam 1981, i.e. BVA, D.Š.] quite different from the model-theoretic argument of the later chapters.

Guyer (1992: 100) noticed that some commentators are committed to the assumption that the views of a great philosopher like Kant must possess a profound unity that can be brought out by a sympathetic interpretation. A different interpretation is defended by Guyer himself and the so called "patchwork" theorists: Kant's greatness lies more in some of his particular analyses and arguments and in his recognition of the complexity of the connections among them than in his pretensions to systematicity. I think that something similar is true of Putnam and his interpreters. *Button* and *Sundell* belong to the camp of *unifiers*, *Sher* is clearly a *patchwork* theorist, *Douven* and *Marino* are less explicit, but probably also accept just a juxtaposition, not an amalgamation of BVA and MTA.

Let me start with Putnam himself. The first chapter of *Reason, Truth and History* is dedicated to the BIV scenario, and model theoretic results are briefly mentioned (Putnam 1981: 7), when he says about the BVA argument: "It first occurred to me when I was thinking about a theorem in modern logic, the 'Skolem-Löwenheim Theorem', and I suddenly saw a connection between this theorem and some arguments in Wittgenstein's *Philosophical Investigations*." The prime locus of Wittgensteinian themes seems to be the private language argument: mental representations are not *magically* connected with what they

represent. On the other hand, when discussing the problem of (anti) realism later in the book, the possibility of a BIV scenario is one of the dividing issues between the camps. According to the perspective of metaphysical realism:

> … the world consists of some fixed totality of mind-independent objects. There is exactly one true and complete description of 'the way the world is'. Truth involves some sort of correspondence relation between words or thought-signs and external things and sets of things. I shall call this perspective the externalist perspective, because its favorite point of view is a God's Eye point of view (Putnam 1981: 49).

On the internalist perspective, defended by Putnam, the question of what objects does the world consist of is a question that it only makes sense to ask within a theory or description. 'Truth', in an internalist view, is some sort of (idealized) rational acceptability. A 'Brain in a Vat World' is then only a *story* and not a possible world at all (Putnam 1981: 50):

> For from whose point of view is the story being told? Evidently not from the point of view of any of the sentient creatures in the world. Nor from the point of view of any observer in another world who interacts with this world; for a 'world' by definition includes everything that interacts in any way with the things it contains. ... . So the supposition that there could be a world in which all sentient beings are Brains in a Vat presupposes from the outset a God's Eye view of truth, or, more accurately, a No Eye view of truth — truth as independent of observers altogether.

For a metaphysical realist the truth of a theory consists in its corresponding to the world as it is *in itself*, so the BIV scenario cannot be dismissed. This establishes an elegant connection between MR and BIV in the form of *modus tollens*, in the version of *Sundell* (Goldberg 2016: 229):

1) If metaphysical realism is true, then pervasive error is a coherent possibility.
2) But pervasive error is not a coherent possibility.
3) So metaphysical realism is false.

The first premise is based on the non-epistemic notion of truth inherent to MR: even an empirically adequate theory—a theory that is predictively accurate and that satisfies any theoretical virtue one may like—may still be false (cf. *Douven* in Goldberg 2016: 175). In Putnam's *earlier* writings the BIV scenario sometimes really figured as an illustration of the possibility of pervasive error. According to MR (Putnam 1977: 485)

> THE WORLD is supposed to be independent of any particular representation we have of it—indeed, it is held that we might be unable to represent THE WORLD correctly at all (e.g., we might all be "brains in a vat", the metaphysical realist tells us).
>
> The most important consequence of metaphysical realism is that truth is supposed to be radically non-epistemic—we might be "brains in a vat" and so the theory that is "ideal" from the point of view of operational utility,

> inner beauty and elegance, "plausibility", simplicity, "conservatism", etc., might be false.

But Putnam (1981) does not justify premise (2) above with the *impossibility* of BIV demonstrated by BVA. The main work of justifying the impossibility of pervasive error is done by MTA, an epistemically ideal theory is guaranteed to be true, according to Putnam. As noted by *Sundell*:

> For an ideal theory to be false, it must be the case that the theory fails to correspond to what the world is like on *the correct interpretation of that theory*. But the MTA shows that there is no way to privilege such an interpretation as correct. The theory is guaranteed to be true on some interpretation, and nothing from inside or outside of the theory can show that that interpretation is the wrong one (Goldberg 2016: 229).

But what I find much more doubtful is that for Sundell "... the anti-realist application of the BVA is the same as the anti-realist application of the MTA. Both arguments attack the coherence of pervasive error" (Goldberg 2016: 234). Putnam's aim in his 1981 was to refute three "solutions" to the puzzle of what it is that determines reference and metaphysical realism is not the main target (cf. De Gaynesford 2011: 579). The main problem is the relation of correspondence on which truth and reference depend for MR. Putnam argues that MR cannot offer a satisfactory account of *determinate* referential relations between the words and the things. If one is in BIV the relation of independent correspondence characteristic for MR is not available, so, given MR commitments, the scenario is paradoxical, a puzzler (Putnam 1981: 51). As he notes in his earlier writings, "Suppose we (and all other sentient beings) are and always were "brains in a vat". Then how does it come about that our word 'vat' refers to *noumenal* vats and not to vats in the image?" (Putnam 1977: 487).

We can agree with *Sher* (Goldberg 2016: 208) that the MTA argument shows that (i) we cannot theoretically determine the reference of our words, and that, as a result, (ii) we must renounce the correspondence theory of truth and robust realism. The BVA argument, on the other hand, shows, that (iii) we cannot truly believe that we are BIVs, and that (iv) Cartesian skepticism is thus undermined. MTA is the main weapon against MR and BVA seems to be a different, *juxtaposed* issue. Sher is also critical with respect to Putnam's results—she thinks that the meta-logical considerations that lead Putnam to conclude (i) are irrelevant to a robust realist/correspondence account of reference (I tend to agree).

The other two "patchwork" theorists, *Douven* and *Marino*, do not have much to say about BVA, but they are also critical with respect to the prospects of MTA. According to *Douven* MTA against realism is based on two assumptions:

(CT) Truth is a matter of correspondence to the facts.
(SN) Semantics is an empirical science like any other.

At the time when the MTA was conceived, it was common to think that a semantics could not be scientifically acceptable if its key concepts could not be accounted for in strictly physicalist terms. But (CT) is no longer the only game in the town, specialists working on truth are nowadays more inclined toward some version of *deflationism*. Douven argues, convincingly, that semantics can be pursued in a scientific spirit without necessarily being part of a reductionist–physicalist research program. Thus MTA is no longer supported (Goldberg, ed. 2016: 189).

Marino discusses the question how does the model-theoretic argument look from the point of view of contemporary *naturalism*. She also stresses that naturalistic forms of disquotationalism diverge from or challenge Putnam's own understanding of reference and truth. Her prime example of a contemporary naturalistic philosopher is "the Second Philosopher", from Maddy (2007). The whole idea of metaphysical realism is somehow misguided from the perspective of modern naturalism and, at least from the contemporary perspective, Putnam seems to be fighting a straw man:

> ... the rejection of metaphysical realism seems significant to Putnam only because of his desire for an account that will, from outside the use of our methods, support and justify those methods—a desire the Second Philosopher does not share (*Marino* in Goldberg 2016: 200).

On the other pole of interpretation the main defender of unification is *Button*. He sees a deep connection between Putnam's thoughts on BIVs, on Skolem's Paradox, and on permutations (also called the "cats and cherries" argument from the *Appendix* in Putnam 1981: 217–218). The last two are based on model-theoretical results but Button unites them all in the form of the BIV-style argument. All types of skepticism—permutation-skepticism (the worry is that our words do not refer as they are intuitively supposed to), skolemism (the worry here is that we cannot tell whether there really are uncountable sets, or merely seem to be[1]) and BIV skepticism are self-refuting when considered as types of *internal* skepticism. Internal skepticism is based on assumptions which we ourselves hold, the skeptic raises an antinomy from within our own worldview. The lynchpin of all of the anti-skeptical arguments is self-refutation, if the skeptical scenario were actual, then we would be unable to articulate this (Goldberg 2016: 153).

Button elegantly develops the template in the form of the BIV-style argument, where the core principle is the principle of *disquotation.* According to Brueckner's original assessment (cf. *Pritchard* and *Ranalli* in Goldberg 2016: 78) one can get the proper anti-skeptical conclusion

[1] Let me note a disturbing typo, the argument against the skolemist is stated as (Goldberg 2016: 143):

> (1S) A smallworlder's word 'countable' applies only to countable (H) sets.
> (2S) My word 'countable' applies only to countable (H) sets.
> (3S) So: I am not a smallworlder.

But surely, (2S) should be "My word 'countable' does not apply only to countable (H) sets."

"It is *not* the case that I am a BIV" from "My utterances of 'I am a BIV' are false" only with the help of the additional *disquotation* principle:

(T) My utterances of 'I am a BIV' are true iff I am a BIV.

This looks question-begging. I am entitled to (T) only if I am entitled to assume that I am a normal human being speaking English rather than a BIV speaking referentially defective vat-English. Since I do not know whether I am speaking English or vat-English, I do not know whether the truth conditions of my utterances of 'I am a BIV' are disquotational ones or not. Still, *Button*, *Ebbs*, *Sundell* and in this collection also *Brueckner* (Goldberg 2016: 21–22), they all defend our knowledge of the semantics of our own language (i.e. our language disquotes and we are entitled to (T)). According to *Ebbs* (Goldberg 2016: 27–36) the goal of the argument is not to show, by strictly a priori methods, that we are *not* always brains in vats. Rather, we always start "relying on already established beliefs and inferences, and applying our best methods for re-evaluating particular beliefs and inferences and arriving at new ones" (Goldberg 2016: 31). The point of the BVA is to transform our understanding of the statement that we are not always brains in vats. If we presuppose substantive beliefs that suffice for minimal competence in the use of the words, we may infer that the disquotational premise (T) is true.

This is still very cautious. In the opening article of the collection *Brueckner* now *defends* Putnam's reasoning in the form of the Simple Argument (Goldberg 2016: 21–22):

(1) If I am a BIV, then my tokens of 'tree' do not refer to trees.
(2) My word 'tree' refers to trees.
(3) So, I am not a BIV.

How does he refute his own earlier criticism? How is (2) justified? Brueckner now claims that whichever language is the one that I am speaking (English or vat-English), my language disquotes. This is licensed by my knowledge of the semantics of my own language (Goldberg 2016: 24).

Button is the most resolute of the three—for him the falsity of disquotation is genuinely *unrepresentable*. He considers the following version of BVA (Goldberg 2016: 135):

(1B) A BIV's word 'brain' does not refer to brains.
(2B) My word 'brain' refers to brains.
(3B) So: I am not a BIV.

Premise (1B) is justified by semantic externalism and premise (2B) by defending disquotation in the mother-tongue. To understand, talk or even just present the BIV scenario, we need to rely on disquotation, so the skeptic cannot even raise doubts about (2B)—"premise (2B) is implicitly required by the BIV skeptic herself in the very *formulation* of her skeptical challenge …, to deny (2B) is self-refuting" (*Button* in Goldberg 2016: 137). Without relying upon disquotation the skeptic cannot even present her worry that everyone is a BIV.

For Button a simple argumentative template, based on self-refutation (as exemplified by the BVA), shows us how to defeat skolemism, permutation-skepticism and BIV skepticism and, in so doing, how to overthrow certain philosophical pictures. The process that unifies MTA and BVA is the following (Goldberg 2016: 153):

Step 1. Isolate a particular philosophical picture.
Step 2. Observe that some skeptical challenge is unanswerable, given this picture.
Step 3. Show that the skepticism in question is actually self-refuting (or reliant on magic).
Step 4. Conclude by rejecting the original picture as incoherent (or reliant on magic).

Let me start by noting that this unifying process is very *general*, it could easily fit, for instance, Berkeley's critique of materialism as a particular philosophical picture (given materialism the skeptical challenge is unanswerable, but skepticism is self-refuting, because in order to conceive of mind-independent objects, we must ourselves be conceiving of them.) A road to *idealism* as is often suspected by Devitt in his comments on Putnam? Not necessarily, the process could perhaps also fit some of Wittgenstein's strategies, the point is, rather, that there need not be any *specific* unity in Putnam's discussions of brains in vats, of Skolem's paradox, and of cats and cherries (that all and only those three arguments fit the procedure diagnosed by Button). My sympathies remain with the *patchwork* theorists but as it is clear from the quotes above, in the late seventies there were several lines of thoughts in Putnam's writings, sometimes separate, sometimes intersecting and Button does a great job in his attempt to provide a unified picture (also in his very elegant and "user-friendly" presentation of skolemism and the permutation argument).

Next, is it really impossible to make sense of the statement that we are not always brains in vats being false? It seems to me that our knowledge of semantic features (disquotation) of our own language cannot be *apriori*—it is an established semantic fact that even in plain vernacular English containing empty names (and perhaps vague expressions) disquotation fails. Suppose we take seriously the idea that sentences uttered by BIVs are neither true nor false, because the preconditions for their having a truth value are not fulfilled. The disquotation scheme for sentences is just the Tarski's schema:

(T) "P" is true if and only if P

If truth-value gaps are admitted, then this principle is no longer valid. Sentences with empty terms ('this dagger' when used by someone under a hallucination), lack the disquotational properties. Yet we still seem to be linguistically competent and possess some level of understanding of our words even if disquotation fails. "Quasi-understanding" perhaps, so that BIV's mental states lacking normal referential properties do not count as real thoughts but "quasi-thoughts" only. Still, BIV's are not

like ants, the scenario makes sense only if they are relevantly similar to us—capable of engaging in cognitive mental activities. In the vat I cannot *really* think "I am a brain in a vat" since I cannot think about real world brains and real world vats. But, as *Folina* (Goldberg 2016: 172) rightly notices, it does not follow that I cannot have thoughts that are epistemically identical to the BIV thought or nearly so. Just recall the discussions about *narrow* content—no matter how different the individual's environment were, the belief would have the same content it actually does. *Horgan, Tienson and Graham*, for instance, defend the notion of narrow phenomenology—according to Cartesian intuitions, as they name them, one intuitively judges that the BIV's mental life exactly matches one's own, the BIV has numerous beliefs, both perceptual and non-perceptual, that exactly match one's own "normal" beliefs (Horgan et al. 2004: 297). Can we really exclude this possibility on the grounds of self-refutation? Contrary to *Ebbs* I find the worry of the question-begging nature of the BVA quite persuasive (Brueckner 1986: 160, quoted by *Ebbs* in Goldberg 2016: 36):

> I can conclude from this [argument] that I am a normal human being rather than a BIV—and thereby lay the skeptical problem to rest—only if I can assume that I mean by "I may be a BIV" what normal human beings mean by it. But I am entitled to that assumption only if I am entitled to assume that I am a normal human being speaking English rather than a BIV speaking vat-English. This must be shown by an anti-skeptical argument, not assumed in advance.

The challenge has now really changed—the original worry was the Cartesian possibility of having massively false beliefs, the "new" skeptical worry is how do we know that our terms refer, that the preconditions of our having real thoughts are fulfilled. Or, in words of *Folina*, our inability to think of or about the exact conditions under which we may be deluded implies that the skeptical thought lacks *specificity*, it does not make it incoherent (Goldberg 2016: 172–173). Similar critical voices are represented by *Pritchard and Ranalli* (Goldberg 2016: 75–89). They provide a list of critiques of the anti-skeptical potential of BVA, ending on a pessimistic note—the BIV hypothesis is simply a template for making vivid what might be our actual epistemic predicament. "*Prima facie* it's hard to see why some of those possible truths [truths we cannot conceive] are not skeptical, representing our epistemic predicament in ways that we cannot conceive" (Goldberg 2016: 89). And *Sher* adds (Goldberg 2016: 225): "… if there are conditions under which BIVs could figure out some things about the world, are we as different from them as Putnam thinks we are? Is it absolutely irrational to entertain the possibility that we are them, that we are at least a little bit like them?"

Let me try to summarize the problem of the relationships between BVA, MTA and MR from the perspective of the BIV scenario. Skepticism was traditionally a road to anti-realism (a total denial of knowledge is difficult to sustain, so the "reality" cannot be something that transcends our cognitive abilities) and externalism, in general, was

supposed to be realistic in spirit. One would therefore expect the anti-skeptical argument such as BVA to support realism, but Putnam is more subtle. According to his intersecting lines of thinking only *internal* realism can deliver the anti-skeptical goods. Metaphysical realism is always in the grip of the "mind the gap" warning: even a rationally optimal or 'ideal' theory of the world could be mistaken. Putnam argues that this is not possible, but his main weapon against MR is the model-theoretic argument. Metaphysical realism commits itself to claim that uniquely determinate referential relations exist between what we say (and think) and the world, and MTA challenges *this* claim. This suggests that we should interpret the BIV scenario as a *referential* puzzle for MR and not as a way of showing that pervasive error is incoherent and in this way opposing the view that a theory which gives every appearance of being true might really be radically false.

BVA, on the other hand, is primarily an anti-skeptical argument, but a Putnam-style refutation of radical skepticism looks like a smallpox vaccine which prevents the severest and the rarest form of smallpox only. The BVA excludes just those bad scenarios "cooked up to be vulnerable to the semantical reply" (Christensen 1993: 302), but one remaining is enough to "kill" your knowledge (DeRose 2000: 128). Even on its own terms Putnam's reasoning remains unconvincing as an antidote for skepticism—most of the vast literature has been critical and my presentation might be biased in this respect since the collection is quite balanced between those who assess the anti-skeptical potential of the argument positively (Brueckner, Ebbs, Button, Sundell) and those who are more doubtful (Pritchard and Ranalli, Folina, Sher). The connections between MTA and BVA might be tenuous (to loose to justify six articles out of fourteen altogether in any case), and perhaps some space should be dedicated to the historical dimension of BIV instead (this type of thought experiment did not start with Putnam in 1981). Still this is an excellent collection of papers provoking and extending discussion in various directions, the long-term life of the *brain in a vat* thought experiment seems to be guaranteed.

References

Brueckner, A. 1986. "Brains in a Vat." *The Journal of Philosophy* 83: 148–16.

Christensen, D. 1993. "Skeptical Problems, Semantical Solutions." *Philosophy and Phenomenological Research* 53: 301–321.

Goldberg, S. C. (ed.). 2016. *The Brain in a Vat*. Cambridge: Cambridge University Press.

De Gaynesford, M. 2011. "Putnam's Model—Theoretic Argument." In Hales, D. (ed.). *A Companion to Relativism*. Oxford: Wiley-Blackwell: 569–587.

DeRose, K. 2000. "How Can We Know that We're Not Brains in Vats?" *The Southern Journal of Philosophy*, Spindel Conference Supplement 38: 121–148.

Descartes, R. 2008. *Meditations on First Philosophy*. New York: Oxford University Press.

Guyer, P. 1992. "Kant's Theory of Freedom by Henry E. Allison." *The Journal of Philosophy* 89: 99–110.

Harman, G. 1973. *Thought*. Princeton: Princeton University Press.

Horgan, T., Tienson, J., Graham, G. 2004. "Phenomenal Intentionality and the Brain in a Vat." In Schanz, R. (ed.). *The Externalist Challenge*. Berlin: Walter de Gruyter.

Maddy, P. 2007. *Second Philosophy: A Naturalistic Method*. Oxford: Oxford University Press.

Miščević, N. 2016. "In Defense of the Twin Earth—The Star Wars Continue." *European Journal of Analytic Philosophy* 12 (2).

Putnam, H. 1977. "Realism and Reason." *Proceedings and Addresses of the American Philosophical Association* 50 (6): 483–498.

Putnam, H. 1981. *Reason, Truth and History*. New York: Cambridge University Press.

Putnam, H. 1992. "Replies." *Philosophical Topics* 20 (1): 347–408.

Putnam, H. 1994. "Comments and Replies." In Clark, P., Hale, B. (eds.). *Reading Putnam*. Oxford: Blackwell: 242–295.

Šuster, D. 2016. "Dreams in a Vat." *European Journal of Analytic Philosophy* 12 (2).

Tallis, R. 2011. *Aping Mankind: Neuromania, Darwinitis and the Misrepresentation of Humanity*. London: Routledge.